



## Strathprints Institutional Repository

Grindrod, Peter and Higham, Desmond J. and Kalna, Gabriela and Spence, Alistair and Stoyanov, Zhivko and Vass, J.K. (2008) *DNA meets the SVD*. Mathematics Today, 44. pp. 80-85.

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (<http://strathprints.strath.ac.uk/>) and the content of this paper for research or study, educational, or not-for-profit purposes without prior permission or charge.

Any correspondence concerning this service should be sent to Strathprints administrator: <mailto:strathprints@strath.ac.uk>



Grindrod, Peter and Higham, Desmond J. and Kalna, Gabriela and Spence, Alistair and Stoyanov, Zhivko and Vass, J.K. (2008) DNA meets the SVD. *Mathematics Today*, 44 . pp. 80-85.

<http://strathprints.strath.ac.uk/15052/>

This is an author produced version of a paper published in *Mathematics Today*, 44 . pp. 80-85. This version has been peer-reviewed but does not include the final publisher proof corrections, published layout or pagination.

Strathprints is designed to allow users to access the research output of the University of Strathclyde. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. You may not engage in further distribution of the material for any profitmaking activities or any commercial gain. You may freely distribute both the url (<http://strathprints.strath.ac.uk>) and the content of this paper for research or study, educational, or not-for-profit purposes without prior permission or charge. You may freely distribute the url (<http://strathprints.strath.ac.uk>) of the Strathprints website.

Any correspondence concerning this service should be sent to The Strathprints Administrator: [eprints@cis.strath.ac.uk](mailto:eprints@cis.strath.ac.uk)

# DNA Meets the SVD

Peter Grindrod<sup>1</sup>, Desmond J. Higham<sup>2</sup>, Gabriela Kalna<sup>2</sup>, Alastair Spence<sup>3</sup>, Zhivko Stoyanov<sup>3</sup>, J. Keith Vass<sup>4</sup>

<sup>1</sup> Department of Mathematics and the Centre for Advanced Computing and Emerging Technologies, University of Reading

<sup>2</sup> Department of Mathematics, University of Strathclyde, Glasgow

<sup>3</sup> Department of Mathematical Sciences, University of Bath

<sup>4</sup> The Beatson Institute for Cancer Research, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD

This paper introduces an important area of computational cell biology where complex, publicly available genomic data is being examined by linear algebra methods, with the aim of revealing biological and medical insights.

## Section 1: What's New?

Since the time of Gregor Mendel, biologists have been attempting to understand how genes determine biological properties. Differences in genes largely explain biological diversity. But in spite of this all humans are recognisably the same due to our control systems that respond to driving forces such as feeding, stress, infection, age, sex and environment. These controls operate at all possible levels, many of which can now be studied using high-throughput technology. Microarrays observe the transfer of information from deoxyribonucleic acid (DNA), containing around 30,000 genes, to messenger ribonucleic acid (mRNA). In this way the state of all these genes can be recorded for individual samples. In terms of the functioning of the cell, genes are important because the mRNA that they create goes on to produce proteins, and proteins are the catalysts of all cells' activities. Maybe 20,000 mRNA signals are responsible for the production of proteins in any single human cell and it is thought that major aspects of development, and disease, can be understood at the gene-mRNA-protein level.

Defects in genes (mutations) can contribute to particular diseases, like cancers, but it is unusual for a single mutation to be found in all cases of one type of cancer. For instance the cancer drug Herceptin, recently in the news, is only affective for the 20% of women with breast cancer who overproduce the protein Her-2. Generally, there are two main reasons to subdivide patients into subgroups based on mRNA profiles: first the groups may respond differently to treatment and second biologists want insight into the disease process.

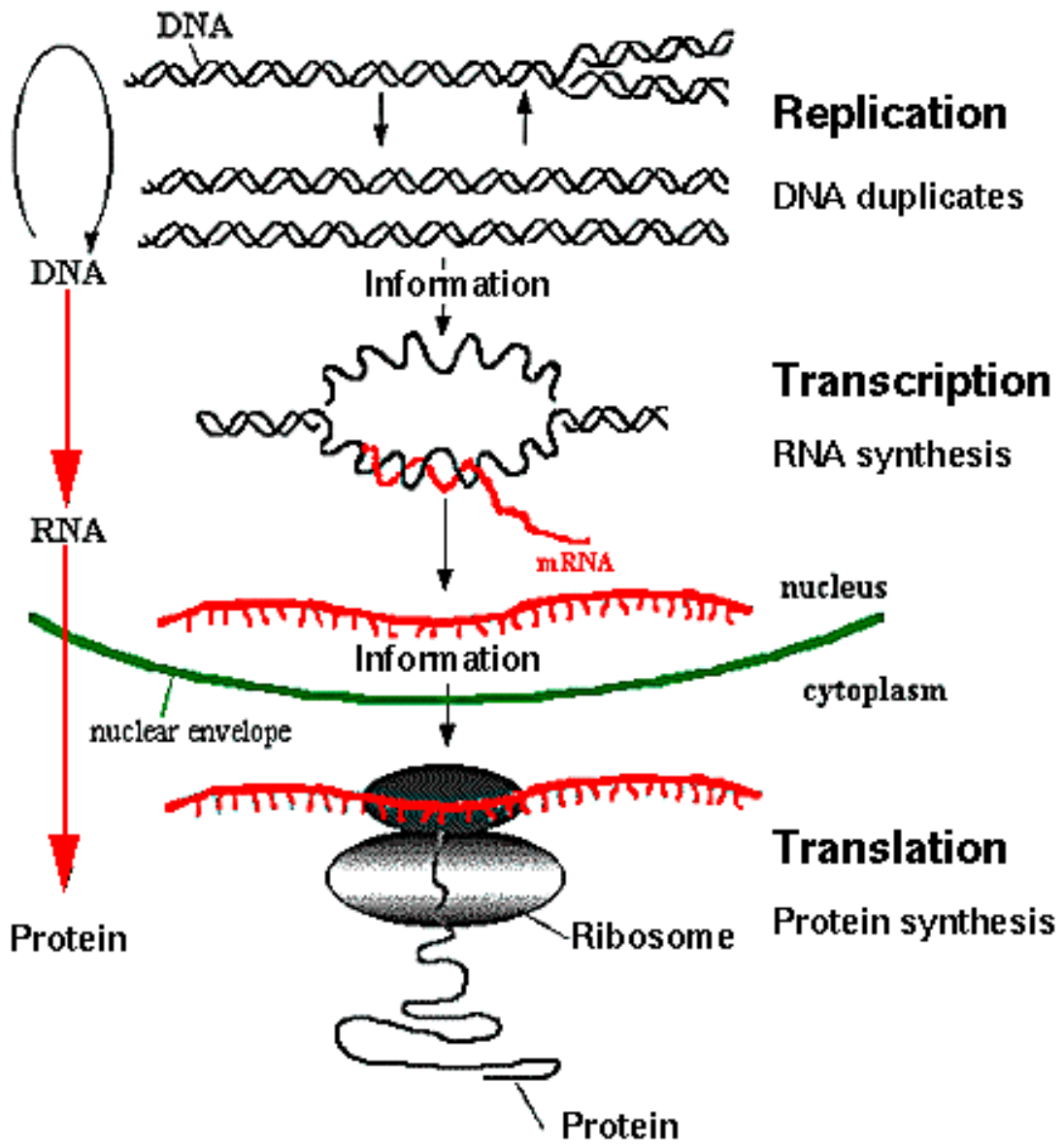
The challenge of making sense of complex genomic information, often involving many genes over relatively few samples, provides many opportunities for mathematicians. Typical data sets take the familiar form of matrices: two dimensional arrays. It is the *size* of the matrices (at least one dimension in the thousands or hundreds of thousands), the level of *uncertainty* in the measurements (repeating experiments to a degree that classical statistical tests are passed is often prohibitively expensive) and the *imprecise* nature of the questions to be addressed, that present the main challenges to applied and computational mathematicians.

Basic biology [Elliot&Elliot03] teaches us that DNA forms an organism's *genetic signature* - arranged as 24 one-dimensional lattices (chromosomes). Analogously, mathematicians know that the singular value decomposition (SVD) of a matrix forms a *spectral signature*, encoding many of its fundamental properties [Golub&VanLoan96]. It is not surprising, therefore, that the SVD is proving to be a valuable tool for teasing meaningful information out of large biological data sets.

Our aim in this article is to give an introductory account of (a) where this type of genomic data comes from and (b) how the SVD can be used to add value. Very little biological and mathematical background is assumed, and the SVD is seen to arise naturally from an algorithmic perspective. Because of its relevance and timeliness, we see this material as ideal for incorporating into undergraduate courses on linear algebra, scientific computing or mathematical biology, or for the basis of an independent study project. References to accessible texts in biology are included in subsequent sections for those who wish to learn more.

## Section 2: What is the Data?

DNA may be viewed as a linear string where each character is one of the four nucleotide bases (C,A,T,G). The string is arranged into a regular double-helix structure. Certain contiguous chunks of DNA, that satisfy known constraints, can code for *genes*. Genes are important because they code for proteins. Proteins are linear strings of amino acids, from an alphabet of 20 characters, but, unlike DNA, these strings fold into complicated 3D shapes, capable of interacting with each other in a myriad of ways. The Central Dogma of Molecular Biology states that a DNA gene specifies its unique mRNA, which in turn specifies its unique protein (Figure 1). This is an oversimplified picture, but it allows useful conclusions to be drawn. There are many references available for those who wish to learn more about basic cell biology from a mathematics/informatics perspective; including [Brazma, Kanehisa03]. We will be concerned with two types of data that give glimpses into the workings of the cell. *Microarrays* are used to estimate simultaneously the amount of each mRNA that is present, so it is now possible to have this information for many thousands of genes in every experiment. At a higher level, *protein-protein interaction* (PPI) data measures which pairs of proteins appear to bind physically, thereby giving clues about the proteins' biological functions.



**Figure 1:** The Central Dogma of Molecular Biology states that a DNA gene specifies its unique mRNA, which in turn specifies its unique protein.

If one gene is responsible for creating one mRNA and hence one type of protein, then each microarray experiment records the activity of each individual gene. This data can be represented by a one-dimensional vector whose  $i$ th entry stores the expression level of the  $i$ th gene. Typically, data from several experiments will be collected. For example, tissue from different cancer patients may be tested, or a single tissue may be tested at different times in order to produce a time series. In both cases, the different experiments are usually referred to as *samples*, and the resulting data set can be thought of as a two-dimensional array, with the  $j$ th column representing the one-dimensional output for the  $j$ th sample. These mRNA measurements are often called

*gene-expression* data and they give a snapshot of the state of transcription of each sample, which in turn reflects the relative importance of their proteins in that tissue. Some leukaemia patients, for example, overproduce Red Blood Cell (RBC) mRNA and proteins including haemoglobin – responsible for the red colour of blood and oxygen transport. In addition to haemoglobin about 40 other RBC genes are “switched-on” at the same time. These proteins are all involved in the architectural and biochemical makeup of RBCs which deliver oxygen from the lungs with every breath and get rid of carbon dioxide at the same time. On the other hand, the normal genes for lung function, including the surfactant protein – which prevent the lungs from collapsing, are often switched-off in lung cancer tumours. The tumour is usually solid and has no “need” for the surfactant. Hence repressed activity levels of these proteins is a possible indicator for that disease.

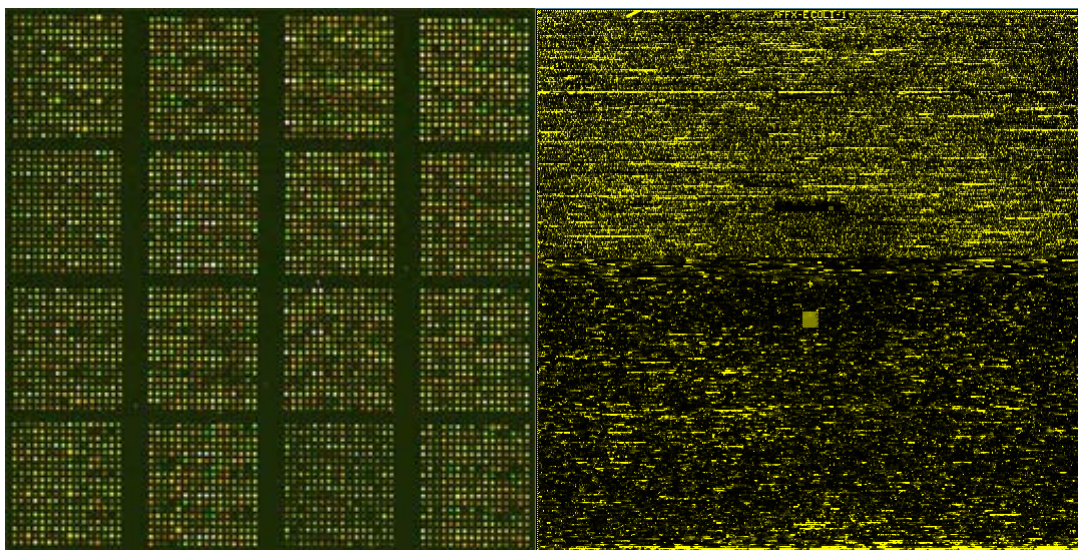
Proteins are, of course, three-dimensional objects, and if two proteins are said to *interact* this means that they physically combine. Experiments can now be conducted where, in principle, every possible pair of proteins in the cell can be tested to see if a mutual interaction takes place. The resulting PPI network is simply an undirected graph whose nodes are proteins and whose edges denote observed interactions [Grindrod04].

In a simplified world where each protein has a single biological function and a protein always forms part of the same biological “team”, the principle of *guilt-by-association* can be effective. The biological function of a protein could be predicted by observing which other proteins had correlated expression levels, or which other proteins were neighbours in the PPI network, assuming that the function of those correlated proteins was known. However, the real story is rarely so clear-cut, with multiple sets of proteins able to perform similar tasks and with single proteins playing multiple roles. So, more flexible, unsupervised, methods are needed to identify the complex networks that maintain the *systems biology* of the cell.

### **Section 3: How is the Data Produced?**

Microarrays depend on complimentary DNA and RNA molecules: this simply means that they form double helices with pairing rules – A matches with T and C with G. A

*sequence* of 25 nucleotides can be synthesized onto about  $10^6$  to  $10^8$  individual positions on an Affymetrix GeneChip© (<http://www.affymetrix.com/index.affx>), while longer pieces of DNA (60 – 100) can be attached to other glass microarrays [Hardiman04]. More unusually now, longer DNA fragments – grown in bacteria, are used in cDNA (copy DNA) microarrays. However, despite production differences, all microarray devices depend on the specific sequence matching rules to ensure that only the required mRNA is detected by either one or several spots (Figure 2). The mRNA being measured is first copied onto a molecule that has some tag, often fluorescent, incorporated – this is what is detected to give a numerical measure of the signal for each spot. The signal has to be evaluated in various ways, depending on the physical design of the device. The casual user is advised to avoid this step and acquire “*normalised*” data, and also to be aware that each step in the preparation of microarray results contributes noise that can affect the reproducibility of the expression data. There are publicly available microarray repositories. Among the most accessible are the National Center for Biotechnology Information ([www.ncbi.nlm.nih.gov/geo](http://www.ncbi.nlm.nih.gov/geo)), the European Molecular Biology Laboratory’s European Bioinformatics Institute ([www.ebi.ac.uk/arrayexpress](http://www.ebi.ac.uk/arrayexpress)), and Stanford MicroArray Database ([genome-www5.stanford.edu/](http://genome-www5.stanford.edu/)) with associated publications and replication of published methods to identify clusters of genes or groups of samples that are useful first steps in beginning to analyse datasets.

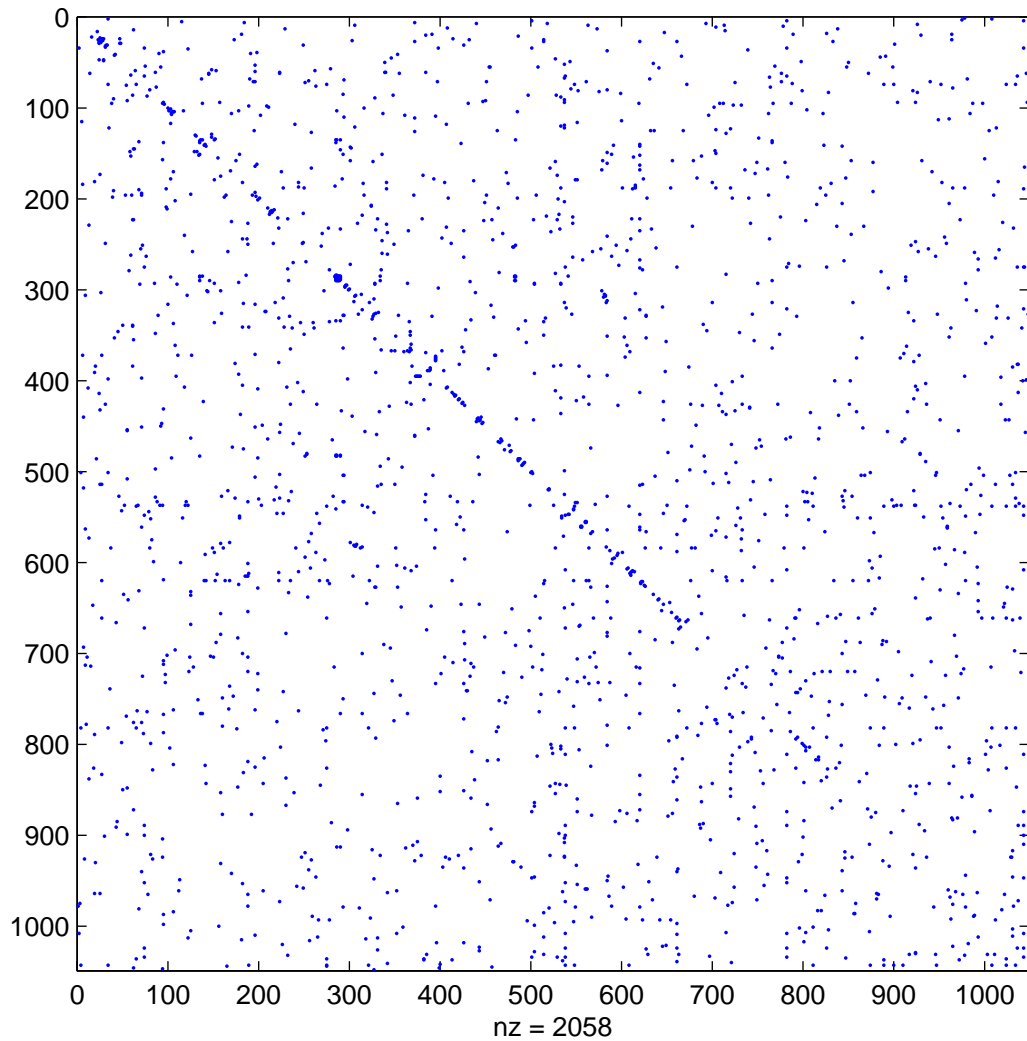


**Figure 2:** c DNA microarray (left) and Affymetrix GeneChip© (right); each gene represented by one or several spots.



To fix notation, we will let the microarray data take the form of a real  $M \times N$  matrix,  $W$ , where  $M$  is the number of genes (rows) and  $N$  is the number of samples (columns), with  $W_{ij} \geq 0$  measuring the activity (expression level) of gene  $i$  in sample  $j$ . The larger the value  $W_{ij}$ , the greater the activity level.

As noted in the previous section, it is also possible to analyse proteins directly and see which proteins can physically interact (bind or become attracted) with one another. Yeast two hybrid (Y2H) experiments allow biologists to measure, in a pair-wise fashion, whether proteins interact. The two hybrid system is based on the premise that many eukaryotic transcriptional activators consist of two physically discrete modular domains. The DNA binding domain of the transcription factor is expressed as a hybrid protein fused to protein X (the "bait"), the activation domain is fused to protein Y (the "prey"). The domains act as independent modules: neither alone can activate transcription. Only if proteins X and Y interact will the activation domain be in the proper position to activate transcription of the reporter gene. As with microarray data, PPI networks obtained this way are very noisy; experimental limitations are believed to result in at least 50% for both the false negative (missing interactions) and false positive (spurious interactions) rates [Grindrod02, Titz04]. In Figure 3 we plot the adjacency matrix of a PPI network for yeast based on the data in [Uetz2000]. Here, a dot in row  $i$  and column  $j$  indicates an interaction between proteins  $i$  and  $j$ . In this case there are 1048 proteins and 1029 interactions.



**Figure 3:** Adjacency matrix for the yeast PPI network from [Uetz2000]. A dot indicates a nonzero.

More accurate, but much less exhaustive, methods for discovering interactions are available, and the data from these can be used to post-process the network producing so-called high-confidence networks [Bader03]

Notwithstanding the inherent uncertainty in the experimental data, we should also keep in mind that the “yes or no”, binary, nature of the PPI network is necessarily an oversimplification. Whether two proteins interact may depend on environmental conditions within the cell and, in particular, on the presence or absence of other proteins. Further, *biological false positives* may be recorded—in this case, two proteins are observed to interact when brought together in the experimental procedure,

but will never have the opportunity to meet in the cell because they operate in distinct physical regions or exist at different times in the cell's life cycle.

In summary, our PPI data takes the form of a real symmetric  $N \times N$  matrix,  $W$ , where  $N$  is the number of proteins, with  $W_{ij} = 1$  if proteins  $i$  and  $j$  are observed to interact and  $W_{ij} = 0$  if there is no interaction.

## Section 4: What are the Questions?

Microarray and PPI data sets provide large-scale, noisy, information. This must be distilled and refined if we are to draw biologically meaningful inferences. The overlapping fields of data mining, dimension reduction and machine learning provide tools for this purpose, and there are already many success stories [Grindrod06]. In this article we focus entirely on one tool—spectral analysis, giving an intuitive, algorithmic derivation of this SVD-based approach.

Typical questions that biologists and medics may ask of high-throughput genomic data include

- 1) Can a disease be linked to a particular set of genes? (For example: Are genes A, B and C almost exclusively overexpressed in patients with disease Y?)
- 2) Given a new tissue sample, can we accurately classify it as either normal or cancerous?
- 3) Given that gene A is known to play an important role in some biological function, can we discover any other genes that behave similarly to gene A and hence may also be involved in this function?
- 4) Can we assign the genes/samples to clusters, where members within each cluster have common behaviour?
- 5) Similarly, can we divide proteins into strongly interconnected clusters?
- 6) Following on from 4) and 5), can we order the genes/samples/proteins so that near-neighbours are highly similar/strongly connected and far-neighbours are very dissimilar/weakly connected?

These issues are clearly interrelated. We will motivate our SVD approach with the ordering problem, 6), although the resulting algorithm also addresses the clustering problem 4)-5) and can be used to tackle questions 1), 2) and 3).

## Section 5: How does the SVD Come in?

We now derive an algorithm, based on biological principles, to tackle the ordering question 6) discussed at the end of the previous section. It turns out that we recover the SVD for a rectangular matrix via the well-known power method to find the dominant eigenvalue of a square matrix.

First, let us repeat the ordering question 6) for samples in a microarray experiment. Imagine that we are trying to find an ordering of the samples on the real line such that close samples exhibit similar gene expression levels, whereas samples far apart on the real line show very different gene expression levels. In order to implement this reordering, our first task is to find a measure with which to compare samples.

As discussed in Section 2, given  $M$  genes and  $N$  samples the gene expression values of the samples are stored in an  $M \times N$  rectangular array  $W$ . We have in mind the case  $M > N$  (many genes and few samples) though, in fact, our ideas and analysis hold equally well for  $M \leq N$ . We shall take as similarity measures the elements in the  $N \times N$  matrix  $W^T W$ . Mathematically, for  $1 \leq i, j \leq N$ , the  $ij$ th element of  $W^T W$  is

$$(W^T W)_{ij} = \sum_{l=1}^M (W^T)_{il} (W)_{lj} = \sum_{l=1}^M W_{li} W_{lj}.$$

Biologically,  $W_{li} W_{lj}$  denotes the product of the expression levels of gene  $l$  in samples  $i$  and  $j$ . So if we sum over all the genes, that is, over  $l = 1, \dots, M$ , we obtain a measure of the total gene expression level for all genes that are expressed in both sample  $i$  and sample  $j$ . We expect this value to be large when sample  $i$  is closely related to sample  $j$ , but small if the samples are unrelated. (Note that all the elements of  $W$  are assumed to be non-negative so no cancellation occurs in summing  $W_{li} W_{lj}$ .)

For convenience we write

$$A = W^T W$$

and assume that  $A$  has row and column sums equal to 1, that is, for row  $i$  and column  $j$ ,

$$\sum_{t=1}^N a_{it} = 1, \quad \sum_{t=1}^N a_{tj} = 1. \quad (1)$$

(This assumption makes the following discussion easier, but it is not essential. If the matrix  $A$  does not have this property a simple rescaling, often called “normalization”, can be implemented, see, for example, [Higham05, Knight06].) We also note that the derivation below applies equally well to the case where proteins are to be ordered from PPI data. In this case we would deal directly with the (symmetric) matrix  $W$ , rather than  $W^T W$ .

Based on the data, our aim now is to assign a real value to each sample in such a way that the ordering of these values reflects a useful ordering of the samples. We start with some initial, arbitrary, set of values and proceed iteratively. Denote the initial position on the real line of the  $i$ th sample as  $x_i^{[0]}$ . We seek an iterative algorithm to reposition the  $i$ th sample based on its relationship with all other samples. We claim that a reasonable candidate for repositioning is

$$x_i^{[k+1]} = \sum_{t=1}^N a_{it} x_t^{[k]} \quad (2)$$

for  $k=0,1,2,\dots$ , with  $k$  counting the number of iterations. In (2), the idea is that the new position of the  $i$ th sample is a weighted combination of the current position of all samples, with the weight for the  $t$ th sample depending on how closely samples  $i$  and  $t$  are related.

However, there is a redundancy in (2), in that all the  $x_i^{[k]}$  could be shifted by an arbitrary amount,  $s$  say, with no change in the ordering, as is seen from

$$\sum_{t=1}^N a_{it} (x_t^{[k]} - s) = \sum_{t=1}^N a_{it} x_t^{[k]} - s \sum_{t=1}^N a_{it} = x_i^{[k+1]} - s \quad (3)$$

using (1).

To remove this redundancy, let us make a shift so that the mean position of the genes is centered at 0. This is implemented as

$$x_i^{[k+1]} = \sum_{t=1}^N a_{it} x_t^{[k]} - \frac{1}{N} \sum_{t=1}^N x_t^{[k]}, \quad (4)$$

that is, we subtract the mean of the  $x_i^{[k]}$  values. Hence

$$\sum_{i=1}^N x_i^{[k+1]} = \sum_{t=1}^N x_t^{[k]} \sum_{i=1}^N a_{it} - \sum_{t=1}^N x_t^{[k]} = \sum_{t=1}^N x_t^{[k]} - \sum_{t=1}^N x_t^{[k]} = 0,$$

using (1). In fact, if the initial ordering has zero mean, that is,

$$\sum_{t=1}^N x_t^{[0]} = 0, \quad (5)$$

then (4) and (2) coincide, ensuring that all future orderings have zero mean, and the freedom expressed by (3) is removed. In practice, round-off errors in evaluating the sums in (4) would cause the mean to drift away from zero, so (4) is less prone to numerical instabilities.

In matrix-vector notation, (4) may be written as

$$\underline{x}^{[k+1]} = \left( A - \frac{1}{N} \underline{1}\underline{1}^T \right) \underline{x}^{[k]} \quad (6)$$

where  $\underline{x}^{[k]}$  is the vector whose  $i^{\text{th}}$  component gives the position of the  $i^{\text{th}}$  sample at the  $k^{\text{th}}$  iteration. Here,  $\underline{1} = (1, 1, \dots, 1)^T$ , and the outer product  $\underline{1}\underline{1}^T$  is the  $N \times N$  matrix with each component equal to 1.

We now make the simple observation that (6) is the well known *power method*, and hence our iterates will converge to an eigenvector [Golub&VanLoan96]. Some straightforward linear algebra now allows us identify this vector and tie it to the SVD.

Because  $A = W^T W$  is a positive semidefinite matrix with non-negative elements, the classical Perron-Frobenius Theorem says that there is an eigenvalue at the spectral radius, and the corresponding eigenvector has strictly positive components [Horn&Johnson85]. Since  $A\underline{x} = \lambda\underline{x} \Rightarrow |\lambda| \leq \|A\|_1 = 1$ , from (1), we know that the spectral radius of  $A$  is less than or equal to 1. But, by construction,  $A\underline{1} = \underline{1}$ , so  $\lambda = 1$  is the dominant eigenvalue of  $A$  with corresponding eigenvector  $\underline{1}$ . In the generic case where  $A$  is irreducible (see, for example, [Horn&Johnson85, p. 361]), this eigenvalue is simple, and the iteration (6) is precisely the power method applied to the matrix

$\left( A - \frac{1}{N} \mathbf{1}\mathbf{1}^T \right)$  with a starting vector satisfying  $\mathbf{1}^T \underline{x}^{[0]} = 0$  (from (5)). We see that the matrix  $A$  has been *deflated*, with the dominant eigenvalue being mapped to zero. Hence the iterates will converge to the eigenvector corresponding to the subdominant eigenvalue (assumed to be simple). In other words, (6) performs a power method iteration that generically converges to a vector corresponding to the second largest eigenvalue of  $A$ .

Finally, we arrive at the SVD. Since  $A = W^T W$ , that subdominant eigenvalue of  $A$  is the square of the second largest singular value of  $W$ , and the converged eigenvector from (6) is the corresponding right singular vector of  $W$ .

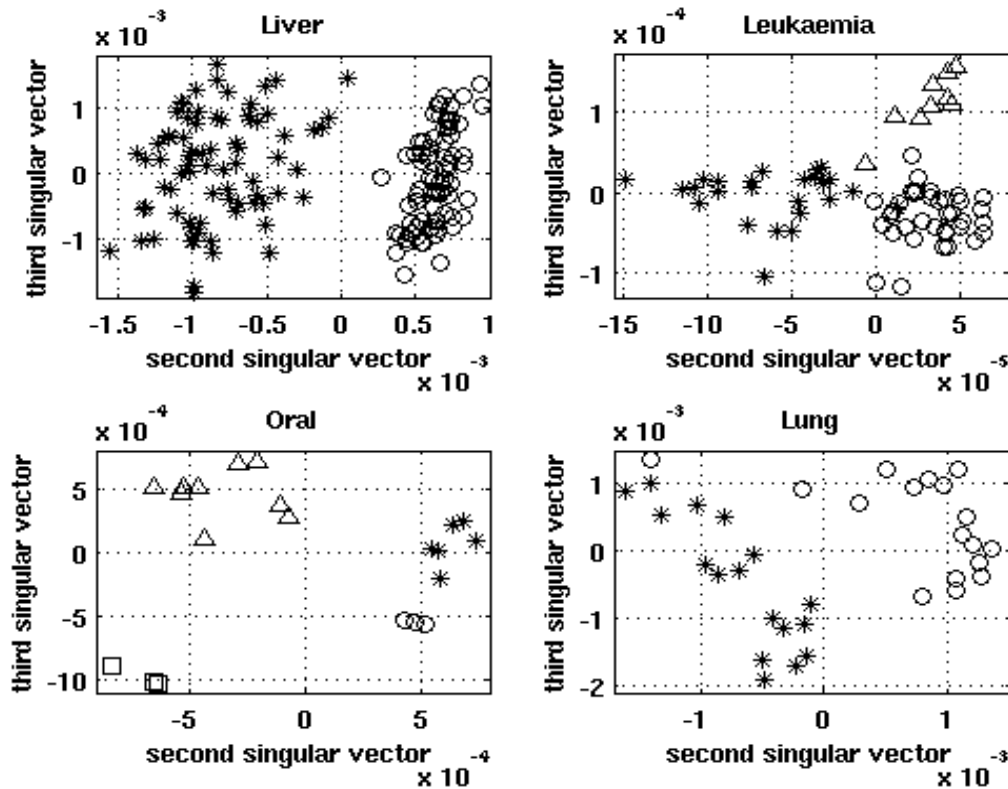
In summary, the ordering arising from (2) with (5), which was motivated by seeking similarities between samples on the basis of connections to genes, yields precisely the second right singular vector of  $W$ . In practice, we need not implement the iteration given by (6). Instead, we could use standard software to compute the second right singular vector of  $W$ , directly obtaining the desired ordering of the samples. Similarly, an algorithm to order genes can be derived as above, but using  $W W^T$ , producing an ordering based on that given by the second left singular vector of  $W$ . Numerical results to show the effectiveness of these ideas are given in the next section.

It is often the case that other re-arrangements of the samples or genes are also relevant, and arguments along the lines of those above can be used to show that the third, fourth, etc. left and right singular vectors are natural candidates. One way to justify this generalisation is developed in [Higham07a], and illustrations also appear in the next section.

## Section 6: Does it Work?

To illustrate the performance of the SVD, we give some results on microarray data from cancer studies. Here, each sample (corresponding to each column of  $W$ ) is from the tumour of a patient with a known type of cancer, or from a normal/control tissue. In these examples, we are simply testing whether the SVD can rediscover the known

groupings, but it should be clear that a successful algorithm has enormous potential for revealing new information and answering questions like those listed at start of section 3.

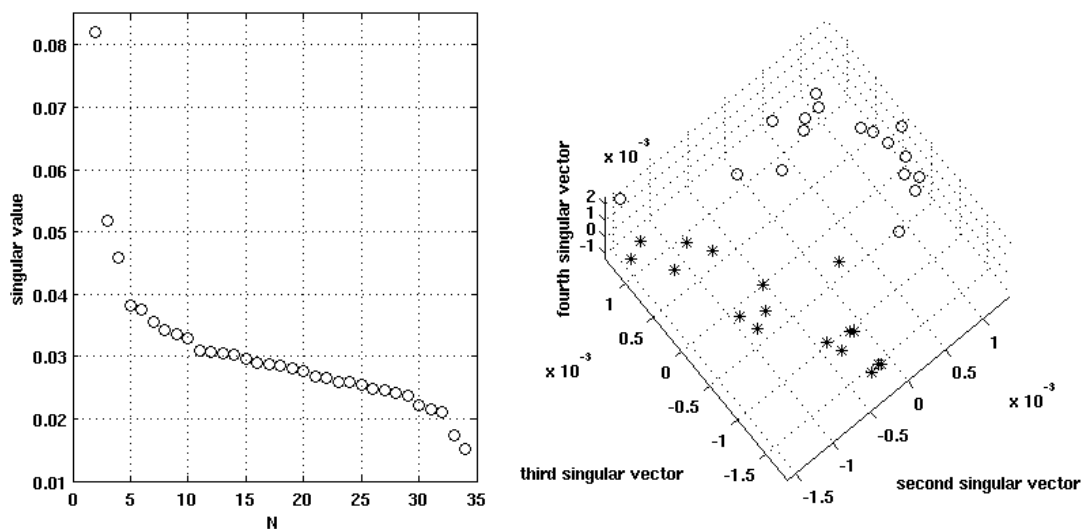


**Figure 4:** Upper left: carcinoma (stars) and normal (circles) liver samples; here  $M=12065$  and  $N=152$ . Upper right: T (triangles) and B (circles) cell ALLs, AML (stars); here  $M=7129$  and  $N=72$ . Lower left: normal in vivo (squares) and in vitro (circles) and carcinomas in vivo (triangles) and in vitro (stars); here  $M=20428$  and  $N=21$ . Lower right: carcinoma (stars) and normal (circles) lung samples; here  $M=5983$  and  $N=34$ .

Figure 4 gives the results. In each of the four cases we have used both the second and third right singular vectors to give two-dimensional components for the samples, producing a “scatter plot” where nearby samples are likely to be related. Let us re-emphasize that the underlying idea here is to use correlation in gene expression behaviour in order to classify sample types. In the upper left scatter plot clear separation of the carcinoma (stars) and normal (circles) kidney samples has been achieved by the second singular vector in data from [Choi05]. The upper right part of Figure 4 shows a scatter plot of the second versus third singular vectors of 72 leukaemia samples from [Golub99]. In this plot the samples are known to divide into

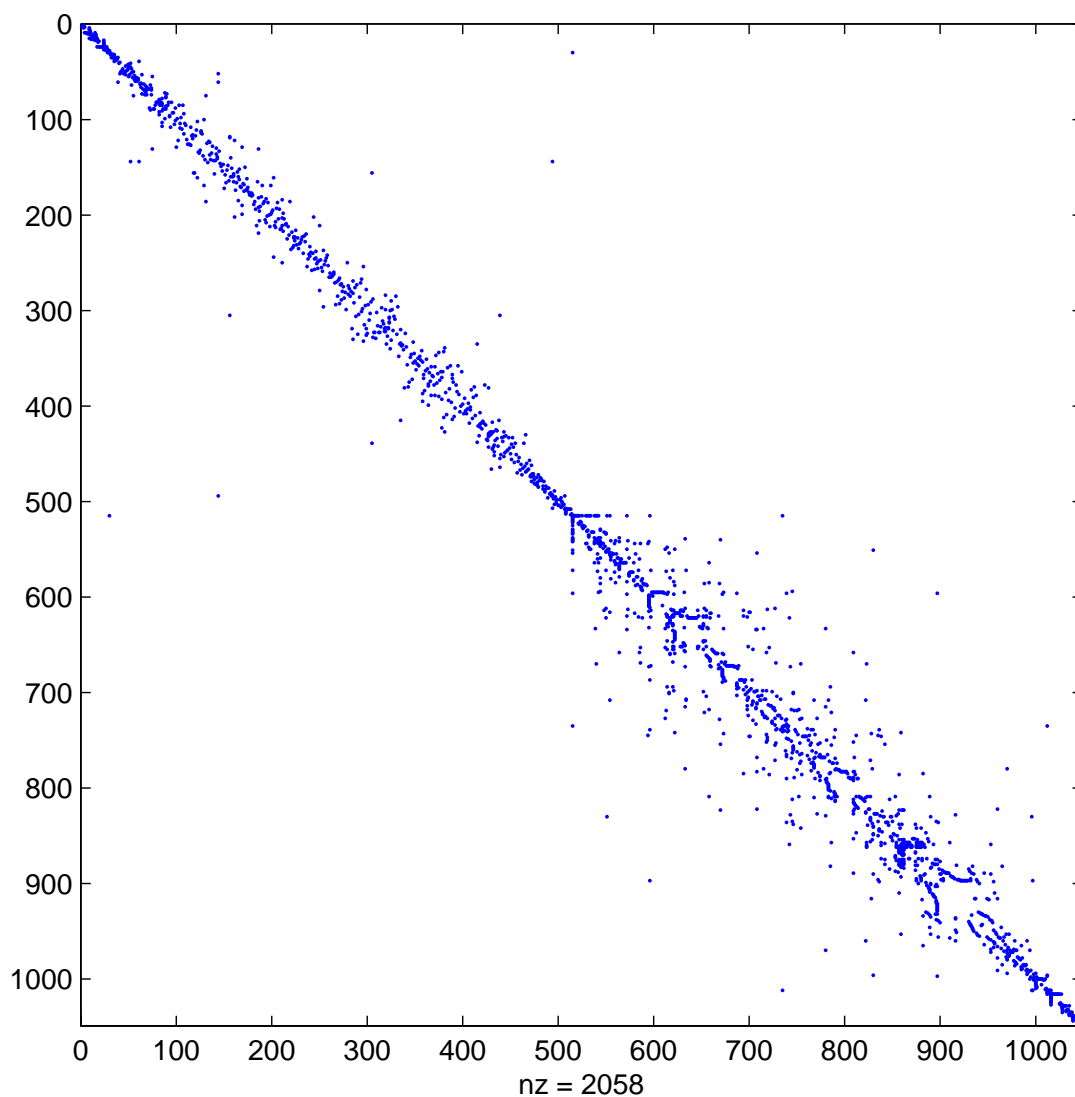


three different groups. We see that the second singular vector does a good job of separating the ALLs and AMLs, while the third singular vector focuses on distinguishing between the T and B subtypes of ALL. Four clear subgroups can be seen in the lower left part of the Figure 4. The second, dominant, singular vector separates *in vivo* and *in vitro* samples and the third vector divides normal samples from oral carcinomas. This data comes from the first set of experiments of an ongoing study led by Dr. Johanna K. Thurlow at the Beatson Institute for Cancer Research, Glasgow. Microarray analysis of normal oral cultures or biopsies and immortal carcinomas grown *in vivo* (xenografts/human tumour) or *in culture* (*in vitro*) has defined a carcinoma gene expression signature which is maintained *in vivo* and *in vitro*. Finally, in the lower right of the figure we can see a scatter plot of two dominant vectors for a set of lung carcinomas and normal lung samples [Choi05]. Although a good separation can be deduced from this plot, it is not performed by a single singular vector. This is an example where carcinomas and normal samples can be distinguished only by using a combination of singular vectors. Although singular value  $\sigma_2$ , shown in Figure 5, left, is far apart from the rest of the singular values the corresponding singular vector alone does not produce clear separation. We see that the next two singular values,  $\sigma_3$  and  $\sigma_4$ , are also separated from the rest of the pack, and an appropriate 3D plot of the lung data set reveals the shapes of two clusters (Figure 5, right).



**Figure 5:** Left: singular values  $\sigma_2, \sigma_3 \dots$  Right: carcinoma (stars) and normal (circles) lung samples.

In Figure 6 we show how the PPI network from Figure 3 looks when it is reordered according to the first subdominant singular vector (also called the *Fiedler vector* in this context). In linear algebra terms, we have applied a symmetric row/column permutation to the adjacency matrix. We see that the new ordering encourages nonzeros to lie close to the diagonal; proteins that appear close together in this new ordering are much more likely to interact than proteins that appear far apart, and the reordering shows that it is possible to view the network as being made up largely of local interactions with relatively few long-range links, as proposed in [Grindrod02].



**Figure 6:** The adjacency matrix for the PPI network in Figure 3 when reordered using the subdominant singular vector (or Fiedler vector).

## Section 6: What Else Can We Do?

Even if we restrict ourselves to the topic of analysing genomic data sets from a computational matrix viewpoint, there are many exciting extensions and variations of the basic ideas presented here. For example,

- Along with clustering/ordering information, it is possible to analyse sensitivity, giving information about how reliable the results are in the presence of uncertainty in the data [Spence07].
- The expression levels in microarray data can take the form of signed data, with negative values representing under-expression [Higham07b].
- Matrix-based inference algorithms can be developed through a systematic Bayesian approach [Lo&Gottardo07], and in the case where data represents evolution, algorithms for time series may be appropriate [Filkov02].
- As an alternative to the SVD, a Non-negative Matrix Factorization can be used [Fogel07].
- Expression data from two different organisms can be analysed with the Generalized Singular Value Decomposition [Alter03].
- Viewing the data as representing a network of interactions, random graph models can be developed [deSilva&Stumpf05] and used to extract biologically meaningful information [Morrison06].
- Generally, in order to add the most value, algorithms that work on multiple sources of data (including discrete quantities from, for example, sequence analysis or interaction and real-valued expression measurements) must be developed [Li06].

Overall, we hope to have convinced the reader that there is a range of important and exciting activities taking place at the data-driven end of computational cell biology where algorithms in matrix computation are being pushed to their limits in the analysis of large, noisy biological datasets.

## **Acknowledgement**

We thank Johanna K. Thurlow (The Beatson Institute for Cancer Research, Glasgow) for useful discussions on the topics in this article.

## **References**

[Alter03] O. Alter, P. O. Brown, D. Botstein,  
Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms, *Proc. Nat. Acad. Sci.* **100** (2003), 3351-3356.

[Bader03] J. S. Bader, A. Chaudhuri, J. M. Rothberg, J. Chant,  
Gaining confidence in high-throughput protein interaction networks, *Nature Biotechnology* **22**, (2003), 78 – 85.

[Brazma] Brazma, A., Parkinson, H. , Schlitt, T. and Shojatalab, M.,  
A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays, European Bioinformatics Institute, on-line tutorial at [http://www.ebi.ac.uk/microarray/biology\\_intro.html](http://www.ebi.ac.uk/microarray/biology_intro.html)

[Choi05] Choi, J.K. and Yu, U. and Yoo, O.J. and Kim, S.,  
Differential coexpression analysis using microarray data and its application to human cancer, *Bioinformatics* **21** (2005), 4348-4355

[Cohen04] Cohen, J.,  
Bioinformatics-An introduction for computer scientists, *ACM Computing Surveys* **36**, (2004), 122-158.

[Elliot&Elliot 03] Elliot, W. H. and Elliot, D. C.,  
Biochemistry and Molecular Biology, Second Edition, Oxford University Press, 2003.

[Filkov02] V. Filkov, S. Skiena, J. Zhi,  
Analysis Techniques for Microarray Time-Series Data, J. Comp. Bio. **9**, (2002), 317-330.

[Fogel07] P. Fogel, S. S. Young, D. M. Hawkins, N. Ledirac,  
Inferential, robust non-negative matrix factorization analysis of microarray data,  
Bioinformatics **23**, (2007), 44-49.

[Golub99] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caliguri, C. D. Bloomfield and E. S. Lander,  
Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science **286** (1999), 531-537.

[Golub&VanLoan96] G. Golub, C. Van Loan (1996), Matrix computations, third edition, The Johns Hopkins University Press, London

[Grindrod02] P. Grindrod,  
Range-dependent random graphs and their application to modeling large small-world Proteome datasets, Phys. Rev. E **66**, (2002), 066702.

[Grindrod04] P. Grindrod, M. Kibble,  
*Review of uses of network and graph theory concepts within proteomics*,  
Expert Review of Proteomics **1**, (2004), 229-238.

[Grindrod06] P. Grindrod, D. J. Higham, J. K. Vass, A. Spence,  
*Systems Biology: unravelling complex networks?*, University of Strathclyde Mathematics Research Report **17**, September 2006.

[Hardiman04] G. Hardiman,  
Microarray platforms - comparisons and contrasts, *Pharmacogenomics* **5** (2004), 487-502.

[Higham05] Higham, D. J. and Kalna, G. and Vass, J. K.  
Analysis of the singular value decomposition as a tool for processing microarray expression data. In: Proceedings of ALGORITMY 2005, 13-18 March 2005, Podbanské, Slovakia.

[Higham07a] Higham, D. J., Kalna, G. and Kibble, M.  
Spectral clustering and its use in bioinformatics, *J. Computational and Applied Math* **204** (2007), 25-37.

[Higham07b] D. J. Higham, G. Kalna, J. K. Vass,  
Spectral analysis of two-signed microarray expression data, *IMA Mathematical+ Medicine and Biology*, 24, **2007**, 131-148.

[Horn&Johnson85] Roger A. Horn and Charles R. Johnson.  
*Matrix Analysis*, Cambridge University Press, 1985.

[Kanehisa03] M. Kanehisa, *Post-genomic Informatics*, Oxford University Press, 2003.

[Knight06] P. A. Knight.  
The Sinkhorn-Knopp algorithm: Convergence and applications.  
Technical Report TR/PA/06/42, 2006, The Parallel Algorithms Project, CERFACS, France.

[Li06] J. Li, X. Li, H. Su, H. Chen, D. W. Galbraith, A framework of integrating gene relations from heterogeneous data sources: an experiment on *Arabidopsis thaliana*., *Bioinformatics* **22**, (2006), 2037-2043.

[Lo&Gottardo07] K. Lo, R. Gottardo,

Flexible empirical Bayes models for differential gene expression  
*Bioinformatics* **23**, (2007), 328 - 335.

[Morrison06] J. L. Morrison, R. Breitling, D. J. Higham and D. R. Gilbert,  
A lock-and-key model for protein-protein interactions, *Bioinformatics* **22**, (2006),  
2012-2019.

[deSilva&Stumpf05] E. de Silva, M. P. H. Stumpf,  
Complex networks and simple models in biology, *Royal Society Interface* **2**, (2005),  
419-430.

[Spence07] A. Spence, Z. Stoyanov, J. K. Vass,  
The sensitivity of spectral clustering applied to gene expression data,  
1<sup>st</sup> International Conference on Bioinformatics and Biomedical Engineering, ICBBE  
2007, July 2007.

[Titz04] B. Titz, M. Schlesner, P. Uetz,  
What do we learn from high-throughput protein interaction data?,  
*Expert Review of Proteomics* **1**, (2004), 111-121.

[Uetz2000]

P.Uetz, L. Giot, G. Cgney, T. A. Mansfield, R. S. Judson, J. R. Knight, E. Lockshon,  
V. Narayan, M. Srinivasan, P. Pochart, A. Qureshi-Emili, Y. Li, B. Godwin,  
D.Conover, T. Kalbfleish, G. Vijayadamodar, M.Yang, M. Johnston, S. Fields, J. M.  
Rothberg,

A comprehensive analysis of protein-protein interactions in *Saccharomyces  
cerevisiae*, *Nature* **403**, (2000), 623—627.